Skeletons Speak Louder than Text: A Motion-Aware Pretraining Paradigm for Video-Based Person Re-Identification

Rifen Lin[1], **Alex Jinpeng Wang**[1], **Jiawei Mo**[1], **Min Li**[1*]

[1]School of Computer Science and Engineering, Central South University
rifen_lin@csu.edu.cn, limin@mail.csu.edu.cn

Abstract

Multimodal pretraining has revolutionized visual understanding, but its impact on video-based person re-identification (ReID) remains underexplored. Existing approaches often rely on video-text pairs, yet suffer from two fundamental limitations: (1) lack of genuine multimodal pretraining, and (2) text poorly captures fine-grained temporal motion—an essential cue for distinguishing identities in video. In this work, we take a bold departure from text-based paradigms by introducing the first skeleton-driven pretraining framework for ReID. To achieve this, we propose Contrastive Skeleton-Image Pretraining for ReID (CSIP-ReID), a novel two-stage method that leverages skeleton sequences as a spatiotemporally informative modality aligned with video frames. In the first stage, we employ contrastive learning to align skeleton and visual features at sequence level. In the second stage, we introduce a dynamic Prototype Fusion Updater (PFU) to refine multimodal identity prototypes, fusing motion and appearance cues. Moreover, we propose a Skeleton Guided Temporal Modeling (SGTM) module that distills temporal cues from skeleton data and integrates them into visual features. Extensive experiments demonstrate that CSIP-ReID achieves new state-of-the-art results on standard video ReID benchmarks (MARS, LS-VID, iLIDS-VID). Moreover, it exhibits strong generalization to skeleton-only ReID tasks (BIWI, IAS), significantly outperforming previous methods. CSIP-ReID pioneers an annotation-free and motion-aware pretraining paradigm for ReID, opening a new frontier in multimodal representation learning.

Introduction

Pretraining has profoundly transformed various areas of computer vision, from image classification (Chen et al. 2020) to multimodal understanding (Li et al. 2021), by learning transferable and robust representations from large-scale unlabeled data. In particular, contrastive pretraining frameworks like CLIP (Radford et al. 2021) have demonstrated remarkable generalization by aligning visual and textual modalities, enabling zero-shot and few-shot capabilities across downstream tasks (Xu et al. 2021; Zhou et al. 2022). Although progress has been made in vision pretraining, video-based person re-identification, which matches individuals across non-overlapping cameras, remains underexplored in terms of cross-modal and pretraining approaches.

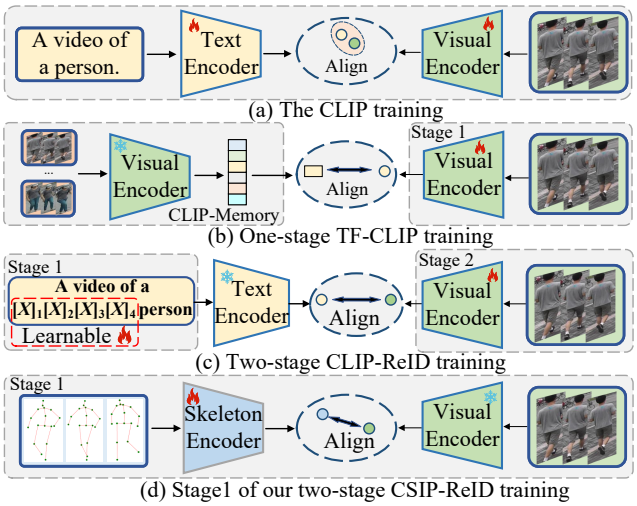

Figure 1: We propose the first contrastive skeleton-image pretraining for ReID. Comparison of CLIP-style learning frameworks: (a) CLIP training. (b) one-stage TF-CLIP training. (c) Two-stage CLIP-ReID training. (d) Contrastive learning in stage 1 of our two-stage CSIP-ReID training.

Some recent studies have attempted to bring CLIP-style frameworks into the ReID domain (Li, Sun, and Li 2023; Yu et al. 2024; Li et al. 2025). However, these methods do not perform genuine multimodal pretraining using ReID datasets. As shown in Fig. 1, they typically reuse the pretrained CLIP visual encoder or text encoder, which is trained on generic image-text pairs that describe objects or scenes rather than person identities. Such text lacks the identity-level semantics necessary for fine-grained discrimination, making the learned modality alignment poorly suited for ReID. Furthermore, these methods neglect the temporal dimension entirely, as their training process does not incorporate motion cues or sequential modeling, which are crucial in video-based re-identification.

To overcome the limitations of prior works, we take a fundamentally different approach. We introduce the first framework that conducts genuine multimodal pretraining directly

*Corresponding author

on ReID datasets, rather than relying on frozen encoders or handcrafted templates derived from unrelated domains. Instead of using text as the second modality, we leverage skeleton sequences—a rich, structured, and annotation-free source of motion information that is naturally aligned with visual inputs in video.

Building on this foundation, we present Contrastive Skeleton-Image Pretraining for ReID (CSIP-ReID), a two-stage framework that learns joint representations from paired skeleton–image sequences. Skeleton offers several key advantages over text. It encodes fine-grained motion patterns that are highly discriminative for person identification, remains robust under appearance or viewpoint variations, and can be efficiently extracted from videos using modern pose estimation models (Goel et al. 2023; Shen et al. 2024). By replacing noisy or generic textual descriptions with expressive motion features, CSIP-ReID establishes a scalable and identity-aware pretraining paradigm tailored for ReID.

After pretraining, we adopt prototype learning for identity supervision, as it effectively aggregates intra-class diversity, including variations in viewpoint and motion, and improves robustness against noisy samples. Specifically, we propose a Prototype Fusion Updater (PFU), which integrates aligned appearance-rich visual features and motion-capturing skeleton features to generate more discriminative and robust prototypes than previous methods (Yu et al. 2024). This is achieved by: (1) Discarding empty frames via skeleton detection; (2) Leveraging background-free skeleton representations to minimize redundancy; (3) Fusing complementary appearance and motion information.

Since the visual encoder lacks spatiotemporal modeling, temporal information across frames is often ignored, causing the task to degenerate into image-based ReID. To address this, we introduce a Skeleton Guided Temporal Modeling (SGTM) module, which captures temporal dynamics through three components: Message Token Encoding (MTE), Auxiliary Temporal Distillation (ATD), and Temporal Aggregation (TA). SGTM distills the strong temporal modeling capability of skeleton as guidance to enhance temporal representation following Learning Using Privileged Information (LUPI) (Vapnik and Vashist 2009) paradigm.

Our main contributions can be summarized as follows:

- We propose CSIP-ReID, the first skeleton-driven pretraining framework for video-based ReID that learns from paired skeleton–image sequences. Unlike prior works that reuse CLIP encoders, our method performs genuine multimodal pretraining on ReID data, establishing a new paradigm beyond text-based approaches.
- We introduce skeletons as a scalable, annotation-free alternative to text for contrastive pretraining. Skeletons are inherently spatiotemporal and identity-discriminative, making them well-suited for motion-aware representation learning in video ReID.
- We design a Prototype Fusion Updater (PFU) using prototype learning to guide visual encoder finetuning and a Skeleton Guided Temporal Modeling (SGTM) module to distill temporal cues from skeletons.
- Our method achieves state-of-the-art performance on both video-based and skeleton-based ReID benchmarks, showcasing its effectiveness and generalization.

Related Work

Video-based Person Re-Identification

Video-based person re-identification aims to extract informative spatial-temporal cues from video sequences to learn robust identity representations. Early works employ CNNs (He et al. 2021b; Liu, Zhang, and Lu 2023) or vision Transformers (Wu et al. 2022; Wang et al. 2025) to capture spatial features. Recently, TF-CLIP (Yu et al. 2024) introduced a CLIP-style approach that replaces the text encoder with a visual memory module. However, it remains unimodal and lacks genuine cross-modal contrastive pretraining. In contrast, our CSIP-ReID performs genuine contrastive pretraining on paired skeleton–image sequences.

For temporal information extraction, existing methods adopt RNNs (Dai et al. 2018), 3D CNNs (Gu et al. 2020), temporal pooling (Wu et al. 2018), attention mechanisms (Liu, Zhang, and Lu 2023) or temporal diffusion (Yu et al. 2024) to capture cross-frame temporal information. Unlike existing methods, we propose a Skeleton Guided Temporal Modeling (SGTM) module that uses skeletons as privileged information to guide temporal feature learning, following the LUPI paradigm (Vapnik and Vashist 2009).

Visual Skeleton Learning

Recent studies have demonstrated the effectiveness of combining skeleton and visual modalities across various tasks. Shao *et al.* (Shao et al. 2021) integrate silhouette image and skeleton features through multimodal fusion. Jiang *et al.* (Jiang et al. 2024) enhance person ReID by guiding visual feature refinement and body-part fusion with skeleton graph modeling. Liu *et al.* (Liu, Chen, and Liu 2024) align visual features and skeleton features via contrastive learning. Lu *et al.* (Lu et al. 2024) transfer high-quality features from X-CLIP to skeleton encoder. 3DAPRL (Jing et al. 2025) leverages 3D pedestrian representations, which are highly similar to skeleton, and introduce shape-aware spatio-temporal modeling to enhance video-based person ReID.

Despite these advances, existing approaches incorporate skeleton by adding separate streams or modules, inevitably increasing model complexity and computational cost. In the era of large-scale models, boosting performance without significantly increasing model size or runtime is essential, techniques such as pretraining (Chen et al. 2023) and knowledge distillation (Xu et al. 2024) offer promising solutions. Inspired by this, CSIP-ReID adopts a skeleton-image contrastive pretraining strategy and distills skeleton information employing prototype learning and LUPI paradigm.

Method

In this section, we present CSIP-ReID, a two-stage framework illustrated in Fig. 2. We describe feature extraction, introduce Stage 1 for contrastive skeleton–image pretraining, and Stage 2 for prototype-guided finetuning with PFU and SGTM modules, followed by the overall training procedure.

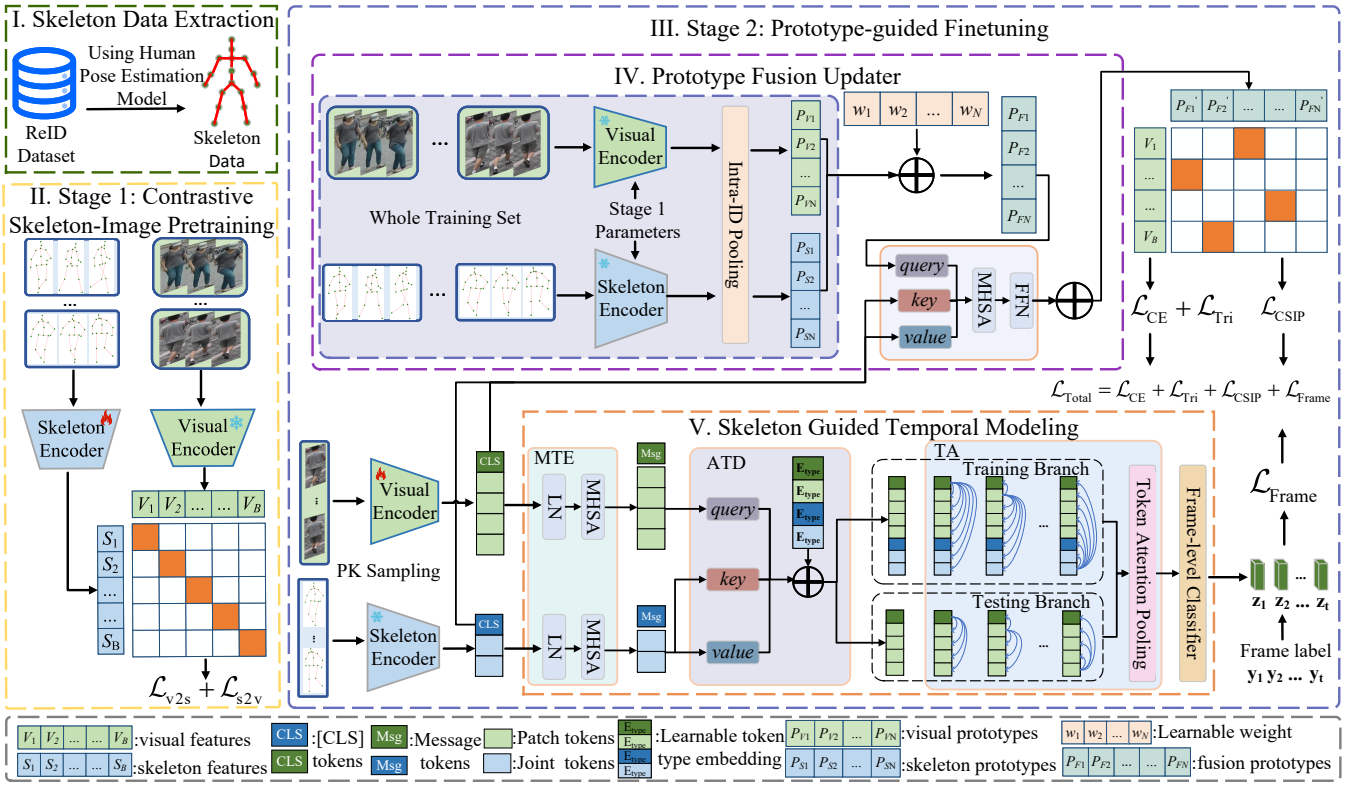

Figure 2: Illustration of the proposed CSIP-ReID framework. (I) Extract skeleton data using human pose estimation model. (II) Stage 1: Contrastive Skeleton-Image Pretraining. (III) Stage 2: Prototype-guided Finetuning, consisting of Prototype Fusion Updater (PFU) and Skeleton Guided Temporal Modeling (SGTM). (IV) Prototype Fusion Updater (PFU) computes and fuses modality-specific prototypes, dynamically updating them with batch visual-skeleton features. (V) Skeleton Guided Temporal Modeling (SGTM) uses MTE to generate message tokens, employs ATD to distill skeleton temporal cues into visual features, and applies TA to aggregate these cues across tokens for frame-level representation.

Feature Extraction with Encoders

CSIP-ReID consists of a visual encoder $\mathcal{V}(\cdot)$ and a skeleton encoder $\mathcal{S}(\cdot)$. We adopt Vision Transformer (ViT) as the visual encoder for its strong spatial modeling capability and proven effectiveness in person re-identification (He et al. 2021a). Meanwhile, Skeleton Graph Transformer (SGT) is employed as the skeleton encoder to capture spatio-temporal patterns from joint graphs, owing to its strong performance in skeleton-based ReID (Rao and Miao 2023).

For the visual modality, an input sequence $\mathbf{V} = \{\mathbf{V}_t\}_{t=1}^{T}$ with frames $\mathbf{V}_t \in \mathbb{R}^{H\times W\times 3}$ is encoded by $\mathcal{V}(\cdot)$, producing frame-level representations $\mathbf{v}_t \in \mathbb{R}^{(1+N_p)\times C}$. Here, T is the number of frames, H and W denote height and width, and N_p is the number of visual patches.

For the skeleton modality, we extract skeleton data from each frame to form $\mathbf{S} = \{\mathbf{S}_t\}_{t=1}^{T}$, where $\mathbf{S}_t \in \mathbb{R}^{J\times 3}$ consists of J joints and 3D coordinates (x, y, z). The skeleton encoder $\mathcal{S}(\cdot)$ processes each frame to produce $\mathbf{s}_t \in \mathbb{R}^{(1+J)\times C}$.

Stage 1: Contrastive Skeleton-Image Pretraining

Stage 1 aims to align skeleton features with those from the frozen CLIP visual encoder through contrastive pretraining, producing well-aligned visual features rich in appearance cues and skeleton features that capture motion.

Specifically, paired skeleton–image sequences are processed by $\mathcal{V}(\cdot)$ and $\mathcal{S}(\cdot)$ to extract frame-level features, which are average-pooled across tokens and frames to obtain sequence-level representations $\bar{\mathbf{v}} \in \mathbb{R}^C$ for the visual modality and $\bar{\mathbf{s}} \in \mathbb{R}^C$ for the skeleton modality. These representations are then used for contrastive pretraining.

By applying this operation to each sample, we obtain a multimodal representation set $\mathcal{M} = \{(\bar{\mathbf{v}}_i, \bar{\mathbf{s}}_i)\}_{i=1}^{N_1}$, where N_1 denotes the number of sequence pairs in stage 1. The similarity between the two modalities is then computed as:

$$\mathrm{Sim}\left(\bar{\mathbf{v}}_i, \bar{\mathbf{s}}_i\right) = \mathcal{J}_v\left(\bar{\mathbf{v}}_i\right) \cdot \mathcal{J}_s\left(\bar{\mathbf{s}}_i\right), \tag{1}$$

where $\mathcal{J}_v$ and $\mathcal{J}_s$ are linear projections into a shared feature space. Similar to CLIP-ReID (Li, Sun, and Li 2023), we adopt $\mathcal{L}_{v2s}$ and $\mathcal{L}_{s2v}$ to align cross-modal features. The visual-to-skeleton contrastive loss $\mathcal{L}_{\text{v2s}}$ is calculated as:

$$\mathcal{L}_{\text{v2s}}(i) = \frac{-1}{|\mathcal{P}_i|} \sum_{j\in\mathcal{P}_i} \log \frac{\exp\left(\mathrm{Sim}(\bar{\mathbf{v}}_i, \bar{\mathbf{s}}_j)/\tau\right)}{\sum_{k=1}^{B} \exp\left(\mathrm{Sim}(\bar{\mathbf{v}}_i, \bar{\mathbf{s}}_k)/\tau\right)}, \tag{2}$$

and the skeleton-to-visual contrastive loss $\mathcal{L}_{\text{s2v}}$:

$$\mathcal{L}_{\text{s2v}}(i) = \frac{-1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \log \frac{\exp\left(\text{Sim}(\bar{\mathbf{v}}_j, \bar{\mathbf{s}}_i)/\tau\right)}{\sum_{k=1}^{B} \exp\left(\text{Sim}(\bar{\mathbf{v}}_k, \bar{\mathbf{s}}_i)/\tau\right)} \quad (3)$$

Here, $\mathcal{P}_i = \{j \mid y_i = y_j\}$ is the set of positive pairs that share the same identity label as the i-th sample. τ is a temperature hyperparameter that controls the sharpness of the distribution. $\mathcal{L}_{\text{v2s}}(i)$ and $\mathcal{L}_{\text{s2v}}(i)$ represent the supervised contrastive loss for aligning visual-to-skeleton and skeleton-to-visual representations, respectively.

Stage 2: Prototype-guided Finetuning

Stage 2 employs prototype-guided finetuning to optimize the visual encoder, as ReID fundamentally relies on its ability to generate discriminative features for identity matching.

Prototype Fusion Updater. As shown in Fig. 2 (IV), the Prototype Fusion Updater (PFU) first combines visual and skeleton modalities to construct fusion prototypes, and then updates them using features within each training batch.

Prototype Fusion. First, we integrate aligned visual and skeleton features to produce more robust and discriminative fusion prototypes. We load the two encoders pretrained during stage 1 and freeze their parameters to ensure consistent feature extraction. Given aligned features $\{\bar{\mathbf{v}}_i\}_{i=1}^{N_1}$ and $\{\bar{\mathbf{s}}_i\}_{i=1}^{N_1}$ with identity labels $\{y_i\}_{i=1}^{N_1}$, we compute modality-specific prototypes by averaging sequence-level features of all samples sharing the same identity, as illustrated by the Intra-ID pooling step in Fig. 2; this step is performed only once during Stage 2.

$$P_{\text{S}}^{(c)} = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \bar{\mathbf{s}}_i, \quad P_{\text{V}}^{(c)} = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \bar{\mathbf{v}}_i, \quad (4)$$

where $\mathcal{I}_c = \{i \mid y_i = c\}$ denotes the set of training samples of identity c. The modality-specific prototypes P_S and $P_V \in \mathbb{R}^{K \times C}$ are then fused by an adaptive fusion mechanism, which learns a dynamic weight $\alpha \in \mathbb{R}^{K \times 1}$:

$$\alpha = \sigma\left(MLP\left([P_S \mid P_V]\right)\right), \quad (5)$$

$$P_F = \alpha P_S + (1 - \alpha) P_V, \quad (6)$$

Here, $\alpha \in \mathbb{R}^{K \times 1}$ is an adaptive weight for each class, learned from concatenated modality features. The symbol $[\cdot \mid \cdot]$ denotes feature-wise concatenation, MLP is a two-layer fully connected network, followed by a sigmoid activation $\sigma(\cdot)$ to constrain the output to $(0, 1)$. This design enables class-aware fusion by dynamically adjusting each modality's contribution, allowing the final prototypes to comprehensively capture both discriminative appearance cues and inherent structural patterns.

Prototype Update. Secondly, we observe that using fixed fusion prototypes limits adaptability as it overlooks the appearance diversity within the same identity. Therefore, we dynamically adjust the fusion prototypes for each input sequence to capture sequence-specific characteristics.

For each training batch, we extract visual features $\mathbf{f}_v$ from the visual encoder and skeleton features $\mathbf{f}_{\text{s}}$ from the skeleton encoder. They are concatenated along the token dimension to form a fused sequence $F \in \mathbb{R}^{B \times (L_{\text{vis}} + L_{\text{ske}}) \times C}$, where B is the batch size, $L_{\text{vis}} = 1 + N_p$ and $L_{\text{ske}} = 1 + J$ are the numbers of visual and skeleton tokens.

This multimodal sequence encodes both appearance and structural cues and serves as the key and value for a cross-attention module. Meanwhile, batch-wise fusion prototypes $P_F \in \mathbb{R}^{B \times K \times C}$, representing K identity prototypes, act as queries. PFU adopts a transformer-style architecture comprising self-attention, cross-attention, and a feed-forward MLP. The update process is formulated as:

$$\hat{P}_F = P_F + \text{MLP}\big(\text{CrossAttn}(\text{SelfAttn}(P_F), F)\big). \quad (7)$$

The update begins with $\text{SelfAttn}(P)$ over prototype tokens to enable inter-prototype interaction. The result then attends to the fused tokens F through cross-attention to capture sample-specific details from multi-modal context. Finally, a feed-forward network $\text{MLP}(\cdot)$ refines the output, which is added back through a residual connection to produce the updated prototypes $\hat{P}_F$.

Prototype Supervision Loss. Then the updated fusion prototypes $\hat{P}_F$ are used to supervise ReID. Given prototypes $\hat{P}_1, \ldots, \hat{P}_K$ and a visual feature f_i from the i-th training sample, the classification loss $\mathcal{L}_{CSIP}$ is defined as:

$$\mathcal{L}_{CSIP}(i) = -\sum_{k=1}^{K} q_k \log \frac{\exp\left(f_i^\top \hat{P}_k\right)}{\sum_{j=1}^{K} \exp\left(f_i^\top \hat{P}_j\right)}. \quad (8)$$

What's more, we follow the strong pipeline (Luo et al. 2019) and adopt both the cross-entropy loss $\mathcal{L}_{\text{CE}}$ with label smoothing and the triplet loss $\mathcal{L}_{\text{Triplet}}$ to jointly optimize the visual encoder.

Skeleton Guided Temporal Modeling. To model temporal dynamics, we propose Skeleton-Guided Temporal Modeling (SGTM) module as shown in Fig. 2 (V).

Message Token Encoding. Given $X^{\text{vis}} \in \mathbb{R}^{T \times (1+N_p) \times C}$, we average all tokens per t to extract a compact message token. Unlike directly using [CLS] token, we adopt average pooling since recent work (He et al. 2022) has shown that patch tokens still retains rich semantic information. Pooled tokens are projected via W_v into a shared space and enhanced by temporal MHSA:

$$\tilde{\mathbf{m}}^{\text{vis}} = \text{MHSA}\big(W_v \, \text{Pool}(X_t^{\text{vis}})\big), \quad (9)$$

where W_v is a learnable linear projection, $\text{Pool}(\cdot)$ denotes average pooling over all tokens at t, and $\text{MHSA}(\cdot)$ refers to temporal self-attention. Skeleton message tokens $\tilde{\mathbf{m}}^{\text{ske}}$ are computed in the same manner.

Auxiliary Temporal Distillation. Following the Learning Using Privileged Information (LUPI) paradigm (Vapnik and Vashist 2009), we leverage skeleton features as privileged information available only during training. ATD employs cross-attention to distill skeleton-guided motion cues into visual message tokens:

$$\hat{\mathbf{m}}^{\text{vis}} = \text{CrossAttn}(\tilde{\mathbf{m}}^{\text{vis}}, \tilde{\mathbf{m}}^{\text{ske}}, \tilde{\mathbf{m}}^{\text{ske}}), \tag{10}$$

To enhance modality awareness, we inject learnable type embeddings $\mathbf{E} \in \mathbb{R}^{4\times C}$ for four token types: $\{\mathbf{x}_{t,i}^{\text{vis}}, \hat{\mathbf{m}}^{\text{vis}}, \mathbf{x}_{t,j}^{\text{ske}}, \tilde{\mathbf{m}}^{\text{ske}}\}$, enabling explicit source differentiation. Such a design has proven effective in (Devlin et al. 2019). By distilling temporal cues from skeleton into visual features under the LUPI framework, ATD strengthens the temporal modeling capacity of the visual stream while keeping inference free from skeleton data.

Temporal Aggregation. TA integrates temporal dependencies across token types by forming a unified sequence $\mathbf{X} \in \mathbb{R}^{L\times BT\times C}$, where L varies with training (which includes all four token types) and messages $\tilde{\mathbf{m}}^{\text{ske}}$) and testing (which includes only visual tokens $\mathbf{x}_{t,i}^{\text{vis}}$ and $\hat{\mathbf{m}}^{\text{vis}}$):

$$\mathbf{X} = \begin{cases} \left[\mathbf{x}_{t,i}^{\text{vis}} \,\|\, \hat{\mathbf{m}}^{\text{vis}} \,\|\, \mathbf{x}_{t,j}^{\text{ske}} \,\|\, \tilde{\mathbf{m}}^{\text{ske}}\right], & \text{if training} \\ \left[\mathbf{x}_{t,i}^{\text{vis}} \,\|\, \hat{\mathbf{m}}^{\text{vis}}\right], & \text{if testing} \end{cases} \tag{11}$$

The unified sequence $\mathbf{X}$ is fed into an attention block as shown in Fig. 2 (V), comprising a temporal self-attention layer followed by a feed-forward network, both equipped with residual connections and layer normalization.

Frame-level Supervision Loss. We apply attention-based pooling over tokens to obtain frame-level logits $\mathbf{z}_{i,t} \in \mathbb{R}^C$, where the attention weights highlight informative tokens and aggregate temporal context into a global representation. The classification loss is then computed as:

$$\mathcal{L}_{\text{Frame}} = -\sum_{i=1}^{B}\sum_{t=1}^{T}\sum_{k=1}^{K} q_{i,t,k} \log p_{i,t,k}, \tag{12}$$

where K is the number of identity classes, $p_{i,t,k}$ is the softmax probability derived from $\mathbf{z}_{i,t}$, and $q_{i,t,k}$ is the corresponding frame-level label. This loss enforces consistent identity predictions across all frames of a sample, enhancing frame-level discriminability and compensating for the reliance on sequence-level features in Stage 1 and PFU.

Traing Strategy

Our training strategy consists of two stages: Contrastive Skeleton-Image Pretraining and Prototype-guided Finetuning. In Stage 1, we load pretrained weights from CLIP, freeze the visual encoder and optimize only the skeleton encoder to align the two modalities via supervised contrastive learning. The training objective is:

$$\mathcal{L}_{\text{stage 1}} = \mathcal{L}_{\text{v2s}} + \mathcal{L}_{\text{s2v}}. \tag{13}$$

In Stage 2, we jointly optimize the visual encoder, Prototype Fusion Updater (PFU), and Skeleton-Guided Temporal Modeling (SGTM). Specifically, we employ the cross-entropy loss $\mathcal{L}_{\text{CE}}$, triplet loss $\mathcal{L}_{\text{Triplet}}$, prototype-guided supervision loss $\mathcal{L}_{\text{CSIP}}$ from PFU, and frame-level supervision loss $\mathcal{L}_{\text{Frame}}$ from SGTM. Two hyperparameters λ_1 and λ_2 control the contribution of the last two terms:

$$\mathcal{L}_{\text{stage 2}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Triplet}} + \lambda_1\mathcal{L}_{\text{CSIP}} + \lambda_2\mathcal{L}_{\text{Frame}}. \tag{14}$$

Experiments

Datasets and Evaluation Protocols

We evaluate CSIP-ReID on three video-based person ReID benchmarks: MARS (Zheng et al. 2016), LS-VID (Li et al. 2019), and iLIDS-VID (Wang et al. 2014). Following common practice, we evaluate model performance using the Cumulative Matching Characteristic (CMC) curve at Rank-k and mean Average Precision (mAP) (Zheng et al. 2015).

Experiment Settings

Our model is implemented in PyTorch and trained on a single NVIDIA Tesla L20 GPU. We sample 8 frames per tracklet, resize them to 256×128, and apply data augmentation as in TF-CLIP (Yu et al. 2024). Stage 1 is trained for 120 epochs with a batch size of 64, while Stage 2 is trained for 80 epochs using PK sampling (Hermans, Beyer, and Leibe 2017) with 4 identities × 4 tracklets. The skeleton encoder parameters follow those in TranSG (Rao and Miao 2023). Code is available in https://github.com/Rifen-Lin/CSIP-ReID.git.

Comparison with State-of-the-arts

We compare our method with state-of-the-art approaches on three video-based person ReID benchmarks, with results shown in Tab. 1, demonstrating its superior performance.

On MARS, CSIP-ReID achieves the best performance with an mAP of 90.4% and Rank-1 accuracy of 94.2%. It surpasses TF-CLIP by 1.0% in mAP and 1.2% in Rank-1, largely because CSIP-ReID leverages skeleton guidance as a complementary second modality. ***On LS-VID***, CSIP-ReID achieves the best Rank-1 accuracy on LS-VID at 92.5%, surpassing TF-CLIP by 2.1%, and ranks second in mAP, slightly behind CLIMB-ReID, likely due to a broader similarity neighborhood that admits a few hard negatives, which can be mitigated by re-ranking. ***On iLIDS-VID***, CSIP-ReID attains the best Rank-1 accuracy on iLIDS-VID at 97.2%, exceeding CLIMB-ReID by 0.5% and TF-CLIP by 2.7%. The Rank-5 accuracy is 98.2%, slightly below CLIMB-ReID, likely because the small scale of this dataset limits contrastive pretraining in Stage1.

Ablation Study

To evaluate the contribution of each component in our model, we conduct ablation studies on the MARS and LS-VID datasets, with results summarized in Table 2. $Model1$ serves as the baseline, where only the CLIP vision encoder is fine-tuned. $Model2$ refers to TF-CLIP without the Sequence-Specific Prompt and Temporal Memory Diffusion modules, where the visual encoder is fine-tuned solely under the guidance of the visual prototype.

Effectiveness of Prototype Fusion. As shown in the first three rows of Table 2, $Model2$ outperforms $Model1$ on MARS by leveraging visual prototypes, while $Model3$ achieves further gains by jointly using visual and skeleton prototypes, confirming their complementarity. Similar trends are observed on LS-VID, demonstrating that fusion prototypes provide more effective guidance for fine-tuning the visual encoder than visual prototypes alone.

Methods	Source	MARS		LS-VID		iLIDS-VID	
		mAP	Rank-1	mAP	Rank-1	Rank-1	Rank-5
STMP (Liu et al. 2019)	AAAI19	72.7	84.4	39.1	56.8	84.3	96.8
M3D (Li, Zhang, and Huang 2019)	AAAI19	74.1	84.4	40.1	57.7	74.0	94.3
GLTR (Li et al. 2019)	ICCV19	78.5	87.0	44.3	63.1	86.0	98.0
TCLNet (Hou et al. 2020)	ECCV20	85.1	89.8	70.3	81.5	86.6	-
MGH (Yan et al. 2020)	CVPR20	85.8	90.0	61.8	79.6	85.6	97.1
BiCnet-TKS (Hou et al. 2021)	CVPR21	86.0	90.2	75.1	84.6	-	-
CTL (Liu et al. 2021)	CVPR21	86.7	91.4	-	-	89.7	97.0
DIL (He et al. 2021b)	ICCV21	87.0	90.8	-	-	92.0	98.0
CAVIT (Wu et al. 2022)	ECCV22	87.2	90.8	79.2	89.2	93.3	98.0
SINet (Bai et al. 2022)	CVPR22	86.2	91.0	79.6	87.4	92.5	-
SDCL (Cao et al. 2023)	CVPR23	86.5	91.1	-	-	-	93.2
TCVIT (Wu et al. 2024)	AAAI24	87.6	91.7	83.1	90.1	94.3	99.3
TF-CLIP (Yu et al. 2024)	AAAI24	89.4	93.0	83.8	90.4	94.5	99.1
CLIMB-ReID (Yu et al. 2025)	AAAI25	89.7	93.3	**85.0**	91.3	96.7	**99.9**
CSIP-ReID(Ours)		**90.4**	**94.2**	84.2	**92.5**	**97.2**	98.2

Table 1: Comparison with typical methods on MARS, LS-VID and iLIDS-VID.

Model	Prototype Fusion	Prototype Update	SGTM	Params(M)	FLOPs(G)	MARS			LS-VID		
						mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
1	×	×	×	86.95	16.98	85.6	90.4	96.4	80.2	87.2	95.5
2	×	×	×	107.15	16.99	88.3	90.6	96.9	80.0	87.9	95.6
3	✓	×	×	109.78	20.34	88.4	92.3	97.8	82.8	91.0	97.1
4	✓	✓	×	118.98	20.49	89.2	92.5	98.0	82.9	91.1	97.1
5	✓	×	✓	125.91	21.13	90.1	93.4	98.0	83.4	91.5	97.2
6	✓	✓	✓	135.11	21.28	**90.4**	**94.2**	**98.3**	**84.2**	**92.5**	**97.3**

Table 2: Comparison of different modules and the computational cost on MARS and LS-VID.

Effectiveness of Prototype Update. As shown in Tab. 2, compared with $Model3$, adding prototype updates of PFU to $Model4$ improves mAP by 0.8% and Rank-1 accuracy by 0.2% on MARS, with similar gains on LS-VID, demonstrating its effectiveness. This improvement likely stems from online prototype updates of PFU, which adapt to each batch and capture fine-grained details overlooked by static fusion prototypes, yielding more discriminative representations.

Effectiveness of SGTM. As shown in Tab. 2, the proposed SGTM significantly improves performance, with $Model5$ achieving gains of 1.7% mAP and 1.1% Rank-1 accuracy over $Model3$ on MARS, and similar improvements on LS-VID. These results highlight SGTM's effectiveness, as it distills additional temporal cues from skeleton to enhance visual temporal modeling.

Comparison of different temporal fusion methods. To assess different temporal aggregation strategies, we compare several fusion methods on MARS, following TF-CLIP (Yu et al. 2024). As shown in Fig. 3(a), SGTM achieves 94.2% Rank-1 accuracy, surpassing the second-best method, Conv1D, by 1.2%. This improvement stems from SGTM's ability to model temporal dynamics in visual frames while distilling complementary cues from skeletons.

The effect of λ_1. In stage 2, the parameter λ_1 controls the weight of the prototype supervision loss $\mathcal{L}_{\text{CSIP}}$, which enhances identity discrimination through joint visual–skeleton representation. As shown in Fig. 3(b), the model achieves optimal Rank-1 accuracy and mAP at $\lambda_1 = 1.0$. A smaller λ_1 weakens prototype supervision, while a larger value causes overfitting to prototypes and distorts the video feature space. Thus, $\lambda_1 = 1.0$ is adopted for balanced cross-modal supervision and optimal performance.

The effect of λ_2. The parameter λ_2 controls the weight of frame-level supervision loss $\mathcal{L}_{\text{Frame}}$, which enforces fine-grained alignment of individual frame features. As shown in Fig. 3(c), performance improves as λ_2 increases, peaking at 1.3, where temporal discrimination is best. Higher values may cause overfitting to frame-specific noise. Thus, we set $\lambda_2 = 1.3$ to balance precision and generalization.

Visualization

t-SNE visualization. To demonstrate the effect of CSIP-ReID, we visualize t-SNE (Van der Maaten and Hinton 2008) distributions of visually similar pedestrians. In Fig. 4(a) and (b), each color represents an identity, with red circles marking two similar ones. Compared with TF-CLIP which exhibits scattered features and outliers, CSIP-ReID forms more compact and separable clusters, reducing intra-person variance and enhancing inter-person separation.

Focus Region Analysis. Finally, we visualize CAM results in Fig. 5, where warmer colors indicate stronger

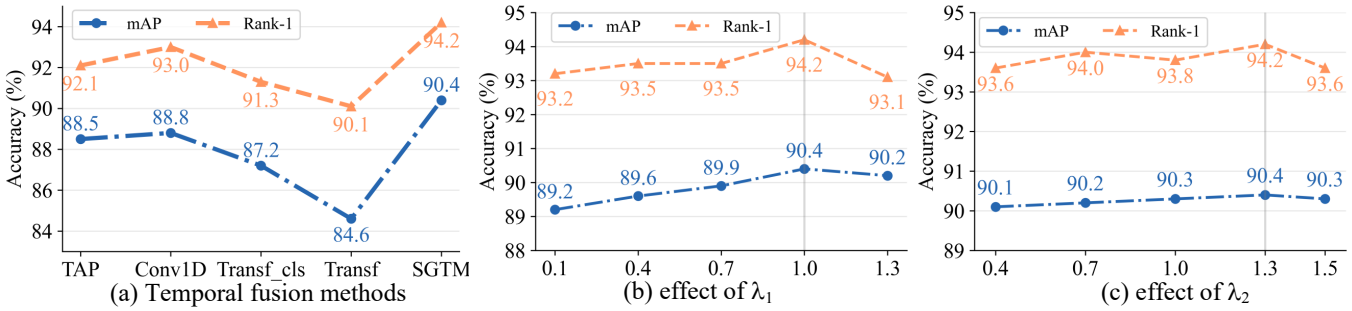

Figure 3: Analysis of key modules/factors affecting performance. This figure illustrates (a) the impact of different temporal fusion methods, (b) the effect of the hyperparameter λ_1, and (c) the effect of λ_2 on model performance.

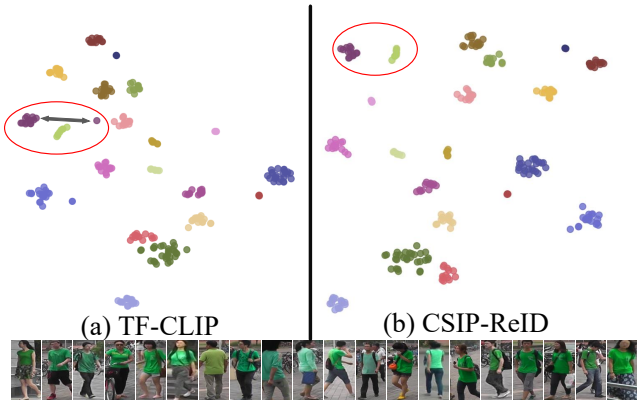

Figure 4: CSIP-ReID produces more compact, discriminative clusters than TF-CLIP in the t-SNE visualization. Each color represents a different identity. Red circles highlight samples from two visually similar identities.

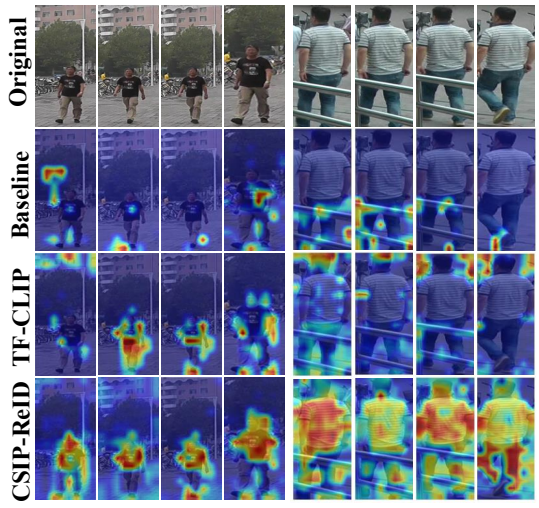

Figure 5: CSIP-ReID shows stronger attention focus on tokens corresponding to human regions. We compare the proposed method CSIP-ReID with Baseline and TF-CLIP.

Methods	BIWI-S		BIWI-W		IAS-A		IAS-B	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1
PoseGait (2020)	9.9	14.0	11.1	8.8	17.5	28.4	20.8	28.9
SGELA (2021)	15.1	25.8	19.0	11.7	13.2	16.7	14.0	22.2
SimMC (2022)	12.3	41.7	19.9	24.5	18.7	44.8	22.9	46.3
Hi-MPC (2024)	17.4	47.5	22.6	27.3	23.2	45.6	25.3	48.2
TranSG (2023)	30.1	68.7	26.9	32.7	32.8	49.2	39.4	59.1
MoCos (2025)	32.1	**72.0**	30.5	36.0	35.8	51.9	45.5	61.5
CSIP-ReID	**34.5**	68.6	**33.8**	**36.9**	**48.1**	**53.6**	**50.7**	**63.3**

Table 3: skeleton-based performance comparison with typical methods on BIWI and IAS.

identity-related attention. The baseline focuses on local cues and fails to capture holistic identity information. TF-CLIP improves attention with visual prototypes and temporal memory but still occasionally attends to background tokens. In contrast, our method leverages skeleton guidance to direct attention primarily toward human regions, yielding more identity-relevant focus.

Transfer to Skeleton-based ReID

To further evaluate the generalization ability of CSIP-ReID, we extend it to skeleton-based ReID by symmetrically modifying the CSIP architecture. The results on the BIWI (Munaro et al. 2014a) and IAS (Munaro et al. 2014b) datasets are shown in Tab. 3. Importantly, visual information is used only during training for cross-modal guidance and excluded during testing to ensure fair comparison, where CSIP-ReID consistently outperforms state-of-the-art approaches, demonstrating strong generalization ability.

Conclusion

In this paper, we explore the potential of skeleton-image pretraining to enhance ReID. Specifically, we propose a novel two-stage framework named CSIP-ReID. Stage 1 aligns visual and skeleton features using supervised contrastive loss while Stage 2 introduces a Prototype Fusion Updater to fuse motion and appearance cues. A Skeleton-Guided Temporal Modeling module distills temporal information from the skeleton modality. Experiments on three benchmarks demonstrate the effectiveness of CSIP-ReID, and its transfer to skeleton-based ReID highlights strong generalization.

Acknowledgments

This work was supported in part by the High Performance Computing Center of Central South University.

References

Bai, S.; Ma, B.; Chang, H.; Huang, R.; and Chen, X. 2022. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7339–7348.

Cao, C.; Fu, X.; Liu, H.; Huang, Y.; Wang, K.; Luo, J.; and Zha, Z.-J. 2023. Event-guided person re-identification via sparse-dense complementary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17990–17999.

Chen, F.-L.; Zhang, D.-Z.; Han, M.-L.; Chen, X.-Y.; Shi, J.; Xu, S.; and Xu, B. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1): 38–56.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Dai, J.; Zhang, P.; Wang, D.; Lu, H.; and Wang, H. 2018. Video person re-identification by temporal residual learning. *IEEE Transactions on Image Processing*, 28(3): 1366–1377.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Goel, S.; Pavlakos, G.; Rajasegaran, J.; Kanazawa, A.; and Malik, J. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *ICCV*.

Gu, X.; Chang, H.; Ma, B.; Zhang, H.; and Chen, X. 2020. Appearance-preserving 3d convolution for video-based person re-identification. In *Proceedings of the European Conference on Computer Vision*, 228–243.

He, J.; Chen, J.-N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; and Wang, C. 2022. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 852–860.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021a. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15013–15022.

He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; and Hua, X.-S. 2021b. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1490–1501.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hou, R.; Chang, H.; Ma, B.; Huang, R.; and Shan, S. 2021. BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014–2023.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2020. Temporal complementary learning for video person re-identification. In *Proceedings of the European Conference on Computer Vision*, 388–405.

Jiang, W.; Zhu, X.; Gao, J.; and Liao, D. 2024. Skeleton-Guided Spatial-Temporal Feature Learning for Video-Based Visible-Infrared Person Re-Identification. *arXiv preprint arXiv:2411.11069*.

Jing, G.; Gao, P.; Lee, Y.; Hu, Y.; and Zhang, H. 2025. 3D-Aided Pedestrian Representation Learning for Video-Based Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, J.; Wang, J.; Tian, Q.; Gao, W.; and Zhang, S. 2019. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 3958–3967.

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8618–8625.

Li, S.; Leng, J.; Kuang, C.; Tan, M.; and Gao, X. 2025. Video-Level Language-Driven Video-Based Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.

Li, S.; Sun, L.; and Li, Q. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 1405–1413.

Liao, R.; Yu, S.; An, W.; and Huang, Y. 2020. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98: 107069.

Liu, J.; Chen, C.; and Liu, M. 2024. Multi-modality co-learning for efficient skeleton-based action recognition. In *Proceedings of the 32nd ACM international conference on multimedia*, 4909–4918.

Liu, J.; Zha, Z.-J.; Wu, W.; Zheng, K.; and Sun, Q. 2021. Spatial-temporal correlation and topology learning for person re-identification in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4370–4379.

Liu, X.; Zhang, P.; and Lu, H. 2023. Video-based Person Re-identification with Long Short-Term Representation Learning. *arXiv preprint arXiv:2308.03703*.

Liu, Y.; Yuan, Z.; Zhou, W.; and Li, H. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8786–8793.

Lu, M.; Yang, S.; Lu, X.; and Liu, J. 2024. Cross-modal contrastive pre-training for few-shot skeleton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9798–9807.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Munaro, M.; Fossati, A.; Basso, A.; Menegatti, E.; and Van Gool, L. 2014a. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*, 161–181. Springer.

Munaro, M.; Ghidoni, S.; Dizmen, D. T.; and Menegatti, E. 2014b. A feature-based approach to people re-identification using skeleton keypoints. In *2014 IEEE international conference on robotics and automation (ICRA)*, 5644–5651. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rao, H.; Leung, C.; and Miao, C. 2024. Hierarchical skeleton meta-prototype contrastive learning with hard skeleton mining for unsupervised person re-identification. *International Journal of Computer Vision*, 132(1): 238–260.

Rao, H.; and Miao, C. 2022. SimMC: Simple Masked Contrastive Learning of Skeleton Representations for Unsupervised Person Re-Identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1290–1297.

Rao, H.; and Miao, C. 2023. TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rao, H.; and Miao, C. 2025. Motif Guided Graph Transformers with Combinatorial Skeleton Prototype Learning for Skeleton-Based Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6703–6712.

Rao, H.; Wang, S.; Hu, X.; Tan, M.; Guo, Y.; Cheng, J.; Liu, X.; and Hu, B. 2021. A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6649–6666.

Shao, W.; You, Z.; Liang, L.; Hu, X.; Li, C.; Wang, W.; and Hu, B. 2021. A multi-modal gait analysis-based detection system of the risk of depression. *IEEE Journal of Biomedical and Health Informatics*, 26(10): 4859–4868.

Shen, Z.; Pi, H.; Xia, Y.; Cen, Z.; Peng, S.; Hu, Z.; Bao, H.; Hu, R.; and Zhou, X. 2024. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Vapnik, V.; and Vashist, A. 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6): 544–557.

Wang, J.; Gao, X.; Niu, S.; Zhao, H.; Feng, G.; and Lin, J. 2025. Learning discriminative features via deep metric learning for video-based person re-identification. *Expert Systems with Applications*, 286: 128123.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person Re-identification by video ranking. In *Proceedings of the European Conference on Computer Vision*, 688–703.

Wu, J.; He, L.; Liu, W.; Yang, Y.; Lei, Z.; Mei, T.; and Li, S. Z. 2022. CAViT: Contextual alignment vision transformer for video object re-identification. In *Proceedings of the European Conference on Computer Vision*, 549–566. Springer.

Wu, P.; Wang, L.; Zhou, S.; Hua, G.; and Sun, C. 2024. Temporal correlation vision transformer for video person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6083–6091.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5177–5186.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; and Zhou, T. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; and Shao, L. 2020. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2899–2908.

Yu, C.; Liu, X.; Wang, Y.; Zhang, P.; and Lu, H. 2024. TF-CLIP: Learning Text-Free CLIP for Video-Based Person Re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7): 6764–6772.

Yu, C.; Liu, X.; Zhu, J.; Wang, Y.; Zhang, P.; and Lu, H. 2025. CLIMB-ReID: A Hybrid CLIP-Mamba Framework for Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9): 9589–9597.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *Proceedings of the European Conference on Computer Vision*, 868–884.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1116–1124.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.